Efficient sampling for Bayesian inference of conjunctive Bayesian networks

Sakoparnig, T ; Beerenwinkel, N

Abstract: Motivation: Cancer development is driven by the accumulation of advantageous mutations and subsequent clonal expansion of cells harbouring these mutations, but the order in which mutations occur remains poorly understood. Advances in genome sequencing and the soon-arriving flood of cancer genome data produced by large cancer sequencing consortia hold the promise to elucidate cancer progression. However, new computational methods are needed to analyse these large datasets. Results: We present a Bayesian inference scheme for Conjunctive Bayesian Networks, a probabilistic graphical model in which mutations accumulate according to partial order constraints and cancer genotypes are observed subject to measurement noise. We develop an efficient MCMC sampling scheme specifically designed to overcome local optima induced by dependency structures. We demonstrate the performance advantage of our sampler over traditional approaches on simulated data and show the advantages of adopting a Bayesian perspective when reanalyzing cancer datasets and comparing our results to previous maximum-likelihood-based approaches. Availability: An R package including the sampler and examples is available at http://www.cbg.ethz.ch/software/bayes-cbn. Contacts: niko.beerenwinkel@bsse.ethz.ch

DOI: https://doi.org/10.1093/bioinformatics/bts433

BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 18 2012, pages 2318–2324
doi:10.1093/bioinformatics/bts433

Systems biology

Advance Access publication July 10, 2012

Efficient sampling for Bayesian inference of conjunctive Bayesian networks

Thomas Sakoparnig[1,2] and Niko Beerenwinkel[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, 4058 Basel and [2]SIB Swiss Institute of Bioinformatics, Supra-University, Basel, Switzerland

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Cancer development is driven by the accumulation of advantageous mutations and subsequent clonal expansion of cells harbouring these mutations, but the order in which mutations occur remains poorly understood. Advances in genome sequencing and the soon-arriving flood of cancer genome data produced by large cancer sequencing consortia hold the promise to elucidate cancer progression. However, new computational methods are needed to analyse these large datasets.

Results: We present a Bayesian inference scheme for Conjunctive Bayesian Networks, a probabilistic graphical model in which mutations accumulate according to partial order constraints and cancer genotypes are observed subject to measurement noise. We develop an efficient MCMC sampling scheme specifically designed to overcome local optima induced by dependency structures. We demonstrate the performance advantage of our sampler over traditional approaches on simulated data and show the advantages of adopting a Bayesian perspective when reanalyzing cancer datasets and comparing our results to previous maximum-likelihood-based approaches.

Availability: An R package including the sampler and examples is available at http://www.cbg.ethz.ch/software/bayes-cbn.

Contacts: niko.beerenwinkel@bsse.ethz.ch

Received on May 21, 2012; revised on June 25, 2012; accepted on July 3, 2012

1 INTRODUCTION

Cancer progression is an evolutionary process characterized by the accumulation of somatic mutations, including single-nucleotide variants, copy number alterations and changes of DNA methylation. Cells with advantageous mutations that confer a proliferative fitness advantage will eventually dominate the cancerous tissue due to clonal expansion.

Mutations in a number of genes are recurrent and it is therefore believed that they are essential for the development of specific cancer types. Mutations of those genes and the functional changes they induce are often referred to as the hallmarks of cancer (Hanahan and Weinberg, 2011). The development and fixation of certain mutations in the tumour cell population seem to depend on the presence of other mutations (Fearon and Vogelstein, 1990), but, in general, the order of occurrence of mutations is poorly understood.

*To whom correspondence should be addressed.

Recent advances in genome sequencing enable large-scale consortia such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) to produce genomic and epigenomic profiles of cancer samples for a medium number of patients on the order of hundreds. For example, The Cancer Genome Atlas Research Network (2011) recently published exome sequencing, copy number variation, gene expression and DNA methylation data for over 300 ovarian cancer patients.

Over the course of the last decade, researchers have applied probabilistic modelling in order to identify the dependency structure of driver mutations for various cancer types. The models include oncogenetic trees (Desper *et al*., 1999, 2000; von Heydebreck *et al*., 2004; Jiang *et al*., 2000; Szabo and Boucher, 2002), mixtures of oncogenetic trees (Beerenwinkel *et al*., 2004; Rahnenführer *et al*., 2005; Yin *et al*., 2006), probabilistic network models (Hjelm *et al*., 2006; Radmacher *et al*., 2001) and Conjunctive Bayesian Networks (CBN) (Beerenwinkel and Sullivant, 2009; Beerenwinkel *et al*., 2007; Gerstung *et al*., 2009). All these models are based on cross-sectional data where genotyping has been performed once on a cancer tissue sample per patient after diagnosis. CBNs jointly model a partial temporal order for the mutation accumulation process and the probabilities of acquiring these mutations based on this cross-sectional data. In contrast to ordinary Bayesian networks, CBNs assign probabilities of zero to genotypes that are not compatible with the partial temporal order modelled by the CBN.

Inference of CBNs is usually done by maximum-likelihood (ML) estimation. Learning the structure of a CBN from observed data is difficult in the presence of observation error and simulated annealing has been used for this task, but assessing the confidence of the estimates is problematic. It has been proposed to derive confidence values from the change in likelihood when removing edges or from bootstrapping the data and refitting the model. However, these approaches have several caveats. (i) There is no *a priori* optimal strategy for parametrization of the simulated annealing scheme; (ii) proper convergence analysis (i.e. stopping) of the simulated annealing algorithm is difficult and (iii) confidence assessment of discrete structure estimates based on bootstrapping is difficult to interpret. Assessing the confidence of the graph component of probabilistic cancer progression models is particularly important, because the structure of the graph is interpreted as the set of possible mutational pathways. The concern about stability has been reinforced by ML-based simulation studies of mixture models of oncogeentic trees that have shown that the structure of these models can be

estimated reliably only for fairly simple structures (Bogojeska *et al.*, 2008).

Adopting a Bayesian perspective, one can get access to the full posterior distribution over all CBN models and hence the uncertainty of their inference and their intrinsic variability. However, application of standard structure sampling schemes is impractical for Bayesian networks as they suffer from very slow mixing and convergence (Giudici and Castelo, 2003; Madigan and York, 1995). Introduction of more sophisticated versions of the basic structure space move types, including a 'reverse edge' move, has been demonstrated to improve mixing and convergence (Grzegorczyk and Husmeier, 2008). However, the 'reverse edge' move cannot be expected to result in the same improvements for CBNs, because CBNs are specialized Bayesian networks with the property that each directed acyclic graph (DAG) defines a unique set of distributions, i.e. a unique equivalence class. For example, unlike for general Bayesian networks, the graphs $A \rightarrow B$ and $B \rightarrow A$ do not define equivalent CBNs. Reversing an edge is therefore a much more severe alteration in a CBN.

In Section 2, we derive the model and describe an MCMC algorithm designed to overcome local optima induced by dependency structures. We validate the sampler on simulated data, compare our proposed algorithm to a more basic standard structure move-based sampler and finally reanalyze real-world cancer datasets in Section 3.

2 METHODS

2.1 Model

Let us consider a set of n driver loci, for example, genes, chromosome arms, CpG islands or other more complex entities such as pathways. We measure the genotype of m tumour samples from different patients. A binary random variable Z_j indicates whether locus j is mutated ($Z_j = 1$) or not ($Z_j = 0$). The binary random vector $Z = (Z_1,\ldots,Z_n)$ encodes the genotype, i.e. the state of all driver loci. Since measuring the mutation state of a gene is an error-prone process due to, for example, measurement noise or erroneous interpretation of genetic changes, we consider the observed genotype of a cancer sample as a separate binary random vector $X = (X_1,\ldots,X_n)$.

We assume that mutations accumulate according to a partial order '$\prec$', where $l \prec k$ means that locus l has to be mutated before a mutation at k can be manifested in a tumour. If l is a direct predecessor of k, then we call this relationship a cover relation. We define the parent set of l as the set of all loci directly preceding l in the partial order. Locus l will mutate with probability θ_l, only if all parents have been mutated before. We say that a genotype Z is compatible with the poset $\prec$ if $(Z_l, Z_k) \neq (0,1)$ for all poset relations $l \prec k$. The exit set of a genotype Z, denoted Exit (Z), is the set of all loci that are not mutated yet but whose parent sets have been fully mutated. The (Discrete Time) Conjunctive Bayesian Network (Beerenwinkel *et al.*, 2007) is defined as

$$\Pr(Z \mid \prec, \theta) = \prod_{\{k\,:\,Z_k = 1\}} \theta_k \prod_{k \in \mathrm{Exit}(Z)} (1 - \theta_k), \tag{1}$$

if Z is compatible with $\prec$ and zero otherwise.

To account for measurement noise or misinterpretation of genetic changes and to avoid the deterministic impact of incompatible observations, we model observation errors by a simple Bernoulli process with parameter ε, the error probability, that is assumed independent and identical across sites. The probability of observing genotype X given the true genotype Z is

$$\Pr(X \mid Z, \varepsilon) = \varepsilon^{d(X,Z)}(1 - \varepsilon)^{n - d(X,Z)}, \tag{2}$$

where $d(X,Z)$ is the Hamming distance between X and Z.

Cancer progression and measurement are assumed to be independent and the marginal probability of X is

$$\Pr(X \mid \prec, \theta, \varepsilon) = \sum_{Z} \Pr(X \mid Z, \varepsilon) \Pr(Z \mid \prec, \theta), \tag{3}$$

where the sum runs over all genotypes compatible with the partial order $\prec$. The marginal likelihood of the m measured genotypes, denoted D, can then be written as

$$\Pr(D \mid \prec, \theta, \varepsilon) = \prod_{X \in D} \sum_{Z} \Pr(X \mid Z, \varepsilon) \Pr(Z \mid \prec, \theta). \tag{4}$$

Since cancer progression and genotype measurement are assumed to be independent, applying Bayes' theorem we obtain

$$\Pr(\prec, \theta, \varepsilon \mid D) \propto \Pr(D \mid \prec, \theta, \varepsilon) \Pr(\prec, \theta) \Pr(\varepsilon) \tag{5}$$

as the joint posterior distribution of model structure (partial order), mutation probabilities and error probability. We further assume the prior independence $\Pr(\prec,\theta) = \Pr(\theta)\Pr(\prec)$ and that mutation probabilities are independent of network structure. Then, the posterior becomes

$$\begin{aligned} \Pr(\prec, \theta, \varepsilon \mid D) \propto \prod_{X \in D} \sum_{Z} \Big[& \Pr(X \mid Z, \varepsilon) \\ & \times \Pr(Z \mid \prec, \theta) \Big] \prod_{k=1}^{n} \Pr(\theta_k) \Pr(\prec) \Pr(\varepsilon). \end{aligned} \tag{6}$$

For θ_k, we choose a non-informative Beta prior with both shape parameters set to 10^{-5}. We use an improper uniform prior for the network structure, $\Pr(\prec) = 1$. The error process parameter ε reflects a trade-off between little to no structure in the case of a too small ε and arbitrary structures in case of a too high ε as most of the genotype variability is explained by the error process. We use $\varepsilon \sim$ Beta (5,30) as error prior.

2.2 Sampler

We adopt a 'random scan Metropolis–Hastings within Gibbs' sampling scheme to sample from the posterior distribution. We use eight different move types for the construction of a hybrid sampler (Tierney, 1994) to explore the joint discrete structure and continuous parameter space of CBNs.

Each move type defines a specific neighbourhood around any point in the state space. All move types except 'relocate theta' and 'reincarnation' are designed such that the neighbourhoods they are considering for any point in the state space are disjoint. As relocate theta and reincarnation are both symmetric move types (see below) their overlap does not compromise the Metropolis–Hastings ratio.

The acceptance probability α of a proposal sample, denoted by '*', is

$$\begin{aligned} \alpha = \min \Bigg\{ 1, \; & \frac{\Pr((\prec, \theta, \varepsilon)^* \mid D)}{\Pr(\prec, \theta, \varepsilon \mid D)} \times \\ & \times \frac{\mathrm{MSP}((\prec, \theta, \varepsilon)^*)}{\mathrm{MSP}((\prec, \theta, \varepsilon))} \frac{\mathrm{TP}((\prec, \theta, \varepsilon) \mid (\prec, \theta, \varepsilon)^*)}{\mathrm{TP}((\prec, \theta, \varepsilon)^* \mid (\prec, \theta, \varepsilon))} \Bigg\}, \end{aligned} \tag{7}$$

where MSP stands for move selection probability. In each iteration, a move type is randomly selected with probability proportional to its MSP, and then a point in the move type neighbourhood of the current sample is selected. The neighbourhoods are equally weighted for all but the reincarnation move, as discussed below. TP stands for the transition probability (density) from the current sample to the proposal sample given that a certain move was selected.

All moves, except the relocate theta and the 'event exchange' move, are combined with a double relocate theta move. In the following paragraphs, the move types are explained in detail.

2.2.1 Relocate theta The mutation probability of a random node is set to a new value which is sampled from a uniform distribution between 0 and 1. As the proposal value does not depend on the current value one can easily see that this move type is symmetric, i.e. $P((\prec,\theta,\varepsilon)^*|(\prec,\theta,\varepsilon)) = P((\prec,\theta,\varepsilon)|(\prec,\theta,\varepsilon)^*)$. Symmetry of move types is desirable as the Hastings factor in Equation (7) becomes one and does not need to be computed. Additionally, the usage of a uniform proposal distribution is advantageous as structural modifications of the CBN usually require a substantial change of the mutation probabilities. The default move selection probability is 0.5.

2.2.2 Relocate epsilon The error probability ε is set to a new value. The new value is sampled from Beta(2,20). As in the relocate theta move, the proposal value does not depend on the current value and hence the move is symmetric. The default move selection probability is 0.1.

2.2.3 New cover relation A random cover relation that which is not resulting in an invalid poset, is inserted (Fig. 1A). Edges that result in an invalid poset are either causing cycles, are redundant edges or render an existing edge redundant. A redundant edge in the poset is one that is present in the transitive closure but not a cover relation. This move type is asymmetric, hence the transition probability and the reverse transition probability which is the transition probability of the corresponding 'delete cover relation' move have to be computed. This is done by complete enumeration. The default move selection probability is 0.2.

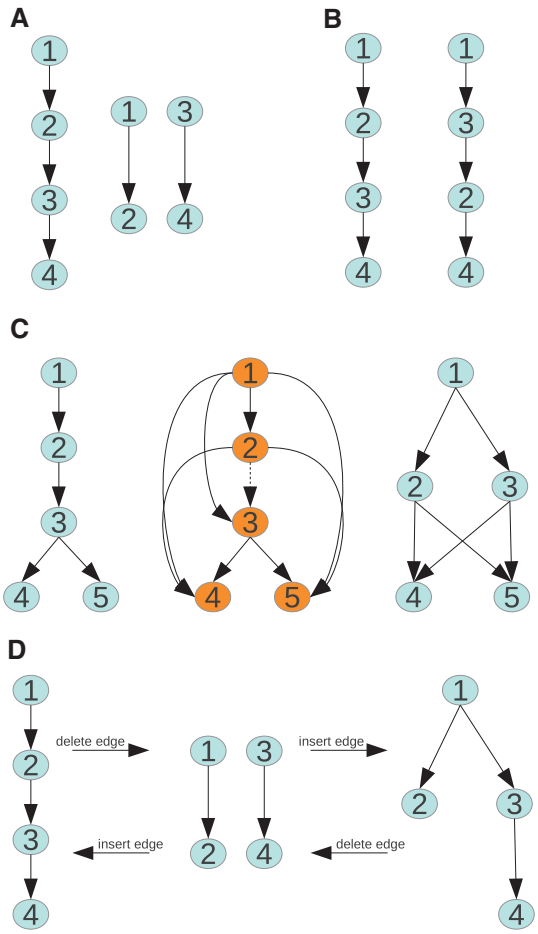

Fig. 1. Move types used for sampling the structure space. (**A**) New/Delete cover relation exemplified between Nodes 2 and 3. (**B**) Event exchange move, performed on Nodes 2 and 3. (**C**) New/delete transitive closure relation. The center graph shows the transitive closure; the dotted edge between Nodes 2 and 3 is the one which is deleted/inserted. (**D**) Reincarnation move. The center graph is the intermediate poset.

2.2.4 Delete cover relation A random edge is removed from the set of cover relations. This move is the asymmetric reverse move to the 'new cover relation' move. The default move selection probability is 0.1.

The moves described so far are standard moves for structure sampling of Bayesian networks and are sufficient for ergodicity. Moves that alter the set of cover relations can induce severe alterations on the according transitive closures. On the other hand, small changes in the transitive closure can require a number of cover relation moves. This is a source of local optima and can compromise convergence and mixing of the MCMC scheme. Therefore, we introduce four new moves types, two of which operate directly on the transitive closure.

2.2.5 New transitive closure relation A random valid edge is inserted into the transitive closure of the poset (Fig. 1C). The set of valid new edges is computed as follows. The transitive closure of the poset is computed, all new edges in the transitive closure that do not trigger additional edges in the transitive closure other than the inserted one and cause at least three changed edges in the corresponding poset are valid. With the last condition, an overlap with the reincarnation move neighbourhood (see below) is avoided. This move type and its reverse counterpart 'delete transitive closure relation' are both asymmetric and their transition probabilities have to be computed. This is again done by enumeration. The default move selection probability is 0.01.

2.2.6 Delete transitive closure relation The set of valid edges to delete is computed as following (Fig. 1C). The transitive closure of the poset is computed, all edges in the transitive closure which can be deleted without corrupting the transitive closure integrity and causing at least three changed edges in the corresponding poset are valid. The default move selection probability is 0.01.

2.2.7 Event exchange The positions of two random nodes in the network topology are exchanged (Fig. 1B). The mutation probabilities of those two nodes are relocated as in the relocate theta move. This move type is symmetric. The default move selection probability is 0.03 .

2.2.8 Reincarnation A delete cover relation move is followed by a new cover relation move (Fig. 1D). This move type is symmetric, which can be seen as follows. The transition probability of this move can be decomposed into the product of the transition probability of the initial delete cover relation move and the following 'new cover relation' move. As both, the current and the final proposal poset have the same number of edges it is easily seen that both delete cover relation moves (proposal and reverse) have the same transition probability. The following 'new cover relation' moves extend the same intermediate poset and hence have an identical neighbourhood and thus identical transition probability. The default move selection probability of the reincarnation move is 0.05.

2.3 Implementation and convergence analysis

We have developed an R package for Bayesian CBNs using the move types described above. The MCMC scheme is implemented in C for performance reasons. As the chains are independent of each other they are ideal candidates for parallelization.

The convergence analysis we use is based on the comparison of samples from multiple chains (Gelman, 2004). This analysis conducts a comparison of the intra- and inter-chain variance for a series of MCMC samples on a parameter-by-parameter basis. It results in a scale-reduction factor $\hat{R}$ for each parameter that reflects the potential reduction of the scale of the current distribution of the parameter of interest if the

simulation was continued indefinitely. Once the maximum $\hat{R}$ over all parameters is <1.1, we sample another round and use the samples derived from this final round for producing summary statistics.

We use methods implemented in the CODA package for convergence analysis of the continuous parameters θ_k and ε and the log-posterior (Plummer *et al.*, 2006). As the mutation probabilities as well as the log-posterior are highly dependent on the underlying dependency structure, we use their convergence as proxies for the convergence of the structure.

3 RESULTS

First, we demonstrate the behaviour of our sampling scheme on simulated data. We then show the performance increase of our sampler over a standard structure move sampler. Finally, we reanalyze real-world cancer datasets in order to demonstrate the increase of interpretability.

3.1 Simulation study

We simulated $N \in \{100,400,800\}$ measured genotypes from one network with 10 nodes and no edges, and another one with 8 edges displayed in Figure 2, assuming an error probability $\varepsilon \in \{0.01,0.1\}$ (Table 1).

Four chains were simulated and monitored for convergence. We thinned out our samples by keeping every 20th sample and produced 25 000 samples per chain and round, i.e. each sampling round consisted of 500 000 iterations per chain. The initial mutation probabilities for each chain were randomly chosen between 0 and 1, the initial error probability was set to 0.05 and the initial poset was set to the empty poset.

Having access to the posterior distribution of the CBNs allows characterization of some properties of these models. Recovery of the structure with confidence first of all depends on the structure itself and the error probability. No structure, i.e. no edges or a high-error probability result in a flat posterior CBN distribution (Fig. 4). Furthermore, the amount of data used for inference is critical (Fig. 4). Nodes that are located in lower parts of the hierarchy and have one or more predecessors with a low mutation probability, also tend to have a flat posterior over their structural dependencies and mutation probabilities. The further down, a node is in the hierarchy, the higher the variance of the posterior mutation probability (Fig. 3A). Uncertainty in the structural environment of a node correlates with uncertainty in the estimation of the mutation probability.

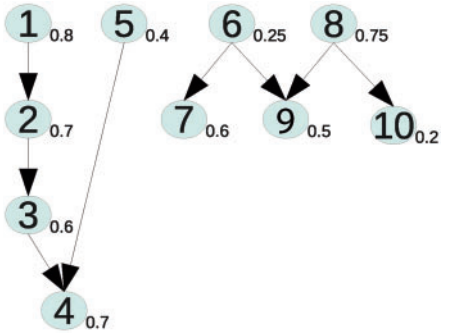

Fig. 2. Poset used for generating data in the simulation study.

The posterior marginal edge probabilities for the dataset, where an empty poset was used for simulation, ranged between 0 and ~0.6 with the bulk being around 0.3 (Fig. 4 bottom right). These numbers are similar for higher genotype numbers N.

In order to obtain quasi-independent samples, further thinning is necessary as autocorrelation is still present. For example, in the $N = 100$ and $\varepsilon = 0.01$ run, the lag-10 values for θ_4 and θ_9 are 0.07 and 0.05, respectively, while the lag-10 values for ε and the log-posterior are 0.67 and 0.69, respectively. The lag-1000 values for ε and the log-posterior are 0.007 and 0.05, respectively. The optimal thinning factor depends on the parameter of interest. Parameters with high-posterior variance show less autocorrelation than parameters with low posterior variance or the structure itself where we use the autocorrelation of the log-posterior (the worst autocorrelation of all quantities) as proxy.

3.2 Performance increase over sampler with standard structure moves

We evaluated the performance increase of the sampling scheme including the new move types new transitive closure relation, delete transitive closure relation, event exchange and reincarnation by comparing it to the basic sampler using only standard structure moves. We changed the move selection probabilities as follows. The relocate theta MSP stays at 0.5, the new cover relation MSP is set to 0.25, the delete cover relation is set to MSP 0.15 and the relocate epsilon MSP stays at 0.1. Everything else is left identical to our proposed sampling scheme.

We simulated data with the same setup as in Section 3.1 and tried to estimate the CBNs without using the new move types. The samplers completed different numbers of rounds within 5 days of running. Only one of the six standard-move runs converged within those 5 days, and this took 12 rounds of sampling. The runs with the newly introduced moves usually converged within two rounds, with the exception of one run, where it took five rounds (Table 1).

3.3 Application to real-world cancer data

We have analysed two cancer datasets. The first one consists of 251 renal cell carcinoma (RCC) cases, which have been analyzed by comparative genome hybridization (CGH), a method that detects chromosomal gains and losses (Jiang *et al.*, 2000). Previous analyses of this dataset by Jiang *et al.* (2000) training

Table 1. Summary of simulation runs for the poset displayed in Figure 2

		Rounds to convergence	
ε	N	New	Standard
0.1	100	1	12
0.1	400	1	>26 (n.c.)
0.1	800	2	>21 (n.c.)
0.01	100	1	>55 (n.c.)
0.01	400	1	>28 (n.c.)
0.01	800	5	>31 (n.c.)

One round of sampling consists of 500 000 iterations per chain. (n.c. = not converged). Details of the run from the fourth row are found in Figure 3.

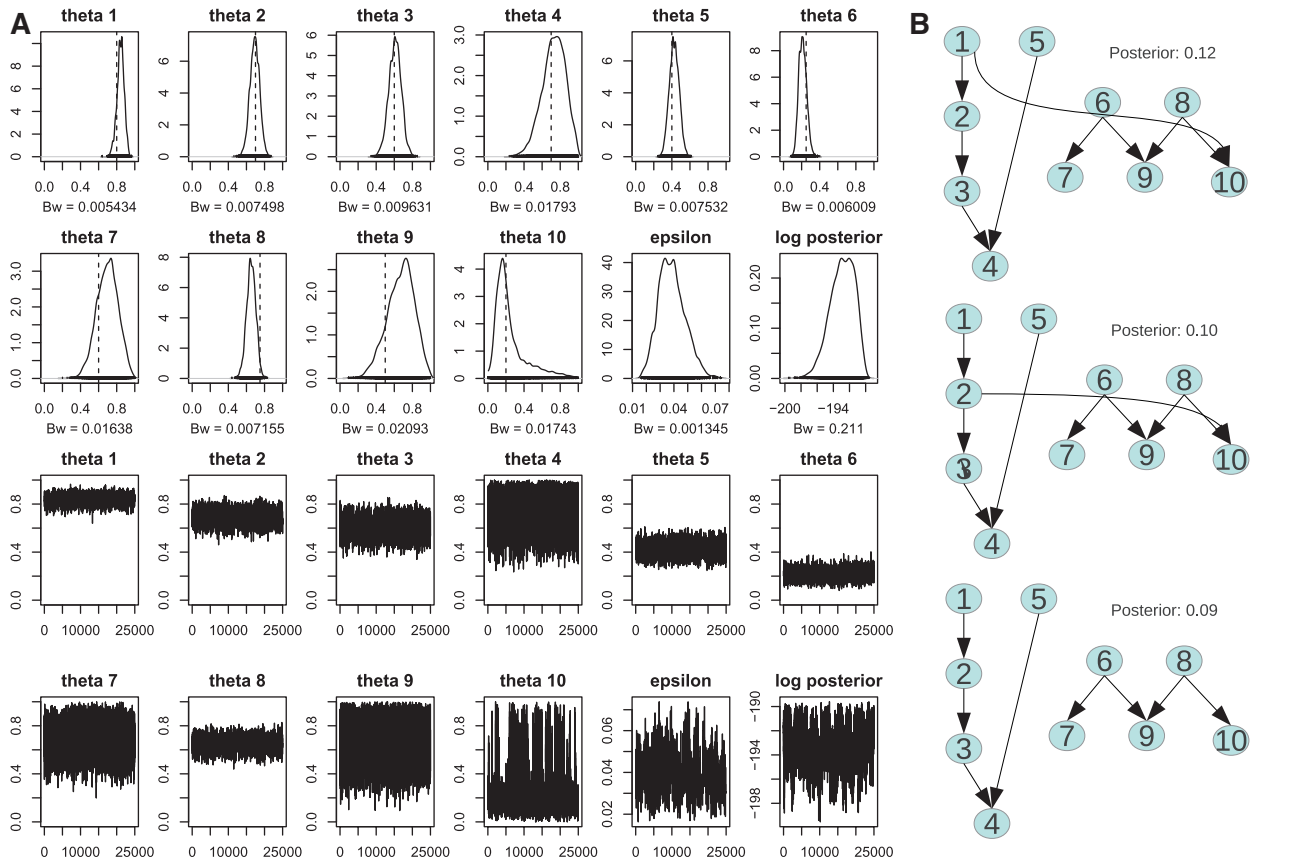

Fig. 3. Summary of sampled CBNs based on simulated data with $N = 100$ cases and error probability $\varepsilon = 0.01$. (**A**) Top two rows present kernel density estimators for all mutation/error probabilities are shown. The dashed bar marks the mutation/error probabilities used for generating the data. The last density plot shows a kernel density estimator for the unnormalized log-posterior. Bottom two rows present trace plots of all mutation, error probabilities and the log-posterior for one of the four chains. Samples used for the trace plots are from the final sampling round. (**B**) The mode of the poset distribution and the posets with the second and third highest frequencies in the MCMC samples are show on the right.

mutagenetic trees and by Gerstung *et al.* (2009) using ML estimation of Continuous Time CBNs resulted in a large disagreement between these two methods. We therefore reanalyzed this dataset and computed the marginal edge probabilities (Fig. 5A). Only three edges had posterior probabilities > 0.5, namely $+17q < +17p$ (posterior 0.82), $-4q < +17q$ (0.73) and $-4q < -6q$ (0.55). Furthermore, it is probable that $-4q$ precedes $-13q$ either directly or indirectly, because the marginal probability of this relation is 0.46, and for the relations $-4q < -6q < -13q$, we found the posterior probabilities 0.55 (as stated above) and 0.41, respectively. All four relationships have been identified by the two previous methods as either direct or indirect. However, both methods claim a number of additional dependencies which we found to have <0.5 posterior probability or they claim direct dependencies, as in the case with $-4q$ and $-13q$, where there might only be an indirect one or vice versa.

The second dataset we analyzed consists of 67 glioblastoma samples (Parsons *et al.*, 2008), in which 16 cancer genes have been DNA sequenced. Following Gerstung *et al.* (2011), we mapped mutated genes measured by sequencing to functional pathways, as defined by Jones *et al.* (2008), to infer order constraints on the level of pathways rather than genes. We did not filter out the secondary type cases as was done by Gerstung *et al.* (2011).

We identified eight cover relations with a posterior marginal edge probability >0.5 (Fig. 5B). Again, for some relations, it is unclear if they are indirect or direct dependencies. Therefore, we computed the marginal probabilities of all relations in the transitive closures of the posets (Fig. 5C). For example, the relation stating that mutation of the TGF-β signalling pathway precedes mutation in the DNA damage control pathway has a posterior probability of 0.7. However, it is probably not a direct relation, but is mediated by either the Hedgehog signalling pathway or the JNK pathway.

4 DISCUSSION

In this work, we have presented a novel Bayesian inference scheme for CBNs, a probabilistic-graphical model of cancer progression that can be estimated from cross-sectional noisy genotype observations. We have developed an efficient MCMC sampling algorithm using a set of moves specifically designed to overcome local optima of dependency structures.

The priors we used in our Bayesian approach are either non-informative or flat. Alternative priors can be used if one wants to introduce more prior biological knowledge, for example, about mutation probabilities, genotype calling errors

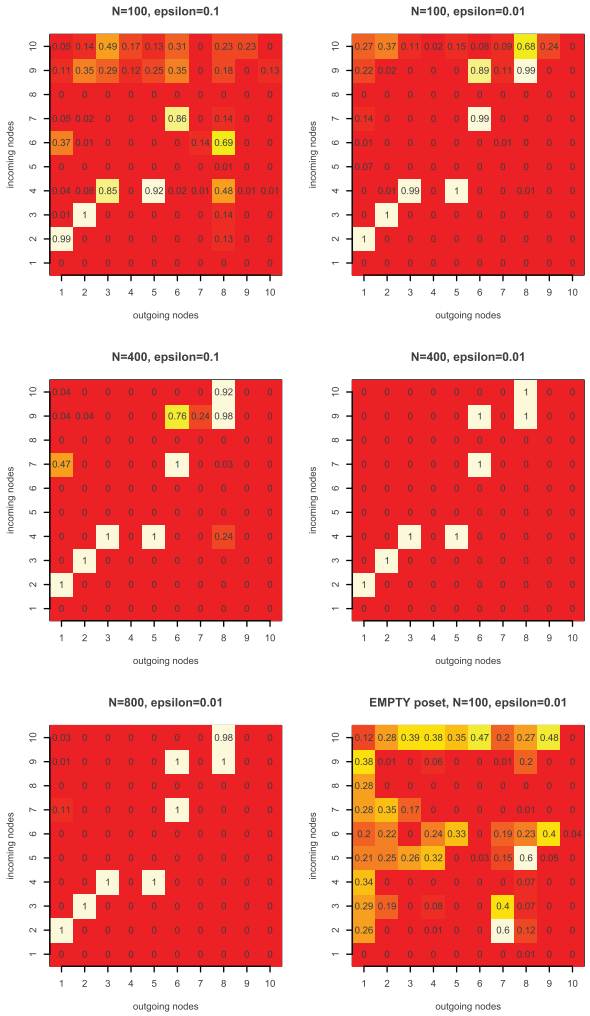

Fig. 4. Marginal cover relation posterior probabilities based on simulated data from various runs. The incoming nodes on the *Y*-axis depend on the outgoing nodes on the *X*-axis.

for specific experimental setups or certain order relations that have been found independently.

For practical purposes, our method is currently capable of handling up to $\sim$15 loci on a standard PC, where the actual run time depends critically on the number of compatible genotypes which are marginalized out Equation (3). Thus, the more closely the partial order dependency structure resembles a linear order, the smaller is the set of compatible genotypes and the faster can the marginal probability of the observed data be computed. In order to reduce the potentially large number of cancer driver mutations, one may apply filtering techniques based, for example, on marginal frequencies, on pairwise correlations or on more sophisticated methods for separating driver from passenger mutations. There is increasing evidence that rather than gene-wise, cancer progression may be more appropriately described on the level of functional pathways. The exact definition of these pathways is ongoing research, but initial studies suggest a small number of cancer-specific pathways on the order of a dozen (Hanahan and Weinberg, 2011; Jones *et al.*, 2008).

The new move types introduced here can be applied to variations of the CBN model, including the Continuous Time CBN (Beerenwinkel and Sullivant, 2009), which models waiting times of mutations and the Isotonic CBN (I-CBN) (Beerenwinkel *et al.*, 2011), which models monotonic progression along a continuous phenotype. We expect that the new move types we added for dealing with local optima of CBNs, especially the moves operating on the transitive closure, may also be useful for other types of Bayesian networks where the directions of all edges can be unambiguously assigned.

Bayesian CBNs are more appropriate for predictive usage than their ML counterparts, since one can reliably assess the confidence of the inferred dependencies as well as the associated mutation probabilities. Drug development and treatment strategies that aim at blocking or hindering cancer progression by targeting certain mutation dependencies may benefit from increased interpretability of confidence assessments. For example, cancer genotypes are used for survival prediction and CBNs have been shown to significantly boost the performance of these predictions (Gerstung *et al.*, 2009; Rahnenführer *et al.*, 2005). The reason

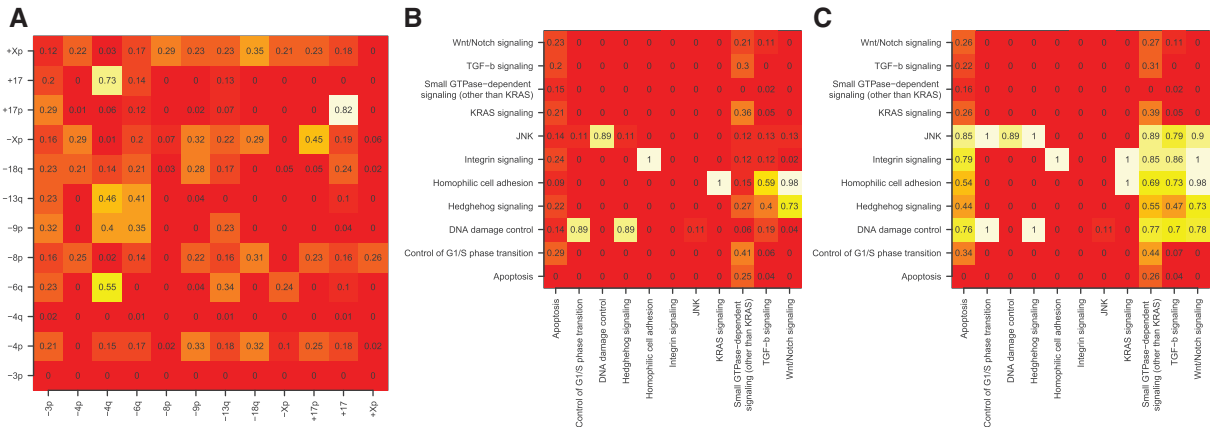

Fig. 5. (**A**) Marginal cover relation posterior probabilities of renal cell carcinoma CGH data. (**B**) Marginal cover relation posterior probabilities of glioblastoma genotypes mapped to functional pathways. (**C**) Marginal posterior probabilities of transitive closure relations from glioblastoma genotypes mapped to functional pathways.

for this improvement is that in predictions for individual patients, strength is borrowed from common features of cancer progression (such as order constraints) that can be learned by CBNs. Using the Bayesian CBN approach introduced here, one can account for model uncertainty in such predictions, for example, using Bayesian estimation of the Cox model.

With international consortia such as TCGA or ICGC producing more and more whole cancer genome screens as well as next-generation sequencing moving into clinical diagnostics, the need for reliable and predictive modelling of genetic progression is growing. We anticipate that our sampling scheme will help to further advance the understanding of cancer progression.

Conflict of Interest: none declared.

REFERENCES

Beerenwinkel,N. and Sullivant,S. (2009) Markov models for accumulating mutations. *Biometrika*, **96**, 1–7.

Beerenwinkel,N. *et al.* (2004) Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.*, **12**, 584–598.

Beerenwinkel,N. *et al.* (2007) Conjunctive Bayesian networks. *Bernoulli*, **13**, 893–909.

Beerenwinkel,N. *et al.* (2011) Learning monotonic genotype-phenotype maps. *Stat. Appl. Genet. Molec. Biol.*, **10**, Article 3.

Bogojeska,J. *et al.* (2008) Stability analysis of mixtures of mutagenetic trees. *BMC Bioinformatics*, **9**, 165.

Desper,R. *et al.* (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.*, **6**, 37–51.

Desper,R. *et al.* (2000) Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.*, **7**, 789–803.

Fearon,E. and Vogelstein,B. (1990) A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.

Gelman,A. (2004) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC, Boca Raton, Florida.

Gerstung,M. *et al.* (2009) Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25**, 2809–2815.

Gerstung,M. *et al.* (2011) The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS One*, **6**, e27136.

Giudici,P. and Castelo,R. (2003) Improving Markov chain Monte Carlo model search for data mining. *Mach. Learn.*, **50**, 127–158.

Grzegorczyk,M. and Husmeier,D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Mach. Learn.*, **71**, 265–305.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hjelm,M. *et al.* (2006) New probabilistic network models and algorithms for oncogenesis. *J. Comput. Biol.*, **13**, 853–865.

Jiang,F. *et al.* (2000) Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. *Cancer Res.*, **60**, 6503–6509.

Jones,S. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.

Madigan,D. and York,J. (1995) Bayesian graphical models for discrete data. *Int. Stat. Rev.*, **63**, 215–232.

Parsons,D.W. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.

Plummer,M. *et al.* (2006) Coda: convergence diagnosis and output analysis for mcmc. *R News*, **6**, 7–11.

Radmacher,M.D. *et al.* (2001) Graph models of oncogenesis with an application to melanoma. *J. Theor. Biol.*, **212**, 535–548.

Rahnenführer,J. *et al.* (2005) Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, **21**, 2438–2446.

Szabo,A. and Boucher,K. (2002) Estimating an oncogenetic tree when false negatives and positives are present. *Math. Biosci.*, **176**, 219–236.

The Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Tierney,L. (1994) Markov chains for exploring posterior distributions. *Ann. Stat.*, **22**, 1701–1728.

von Heydebreck,A. *et al.* (2004) Maximum likelihood estimation of oncogenetic tree models. *Biostatistics*, **5**, 545–556.

Yin,J. *et al.* (2006) Model selection for mixtures of mutagenetic trees. *Stat. Appl. Genet. Molec. Biol.*, **5**, Article17.